REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*
31-12-03

2. REPORT DATE
final report

3. DATES COVERED *(From - To)*
01-02-2002-30-09-2003

4. TITLE AND SUBTITLE
Consolidating the Results of the
CIRCSIM-Tutor Project and Further
Consolidation of the Results
of the CIRCSIM-Tutor Project

5a. CONTRACT NUMBER

5b. GRANT NUMBER
N00014-02-1-0442

5c. PROGRAM ELEMENT NUMBER

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

6. AUTHOR(S)
Evens, Martha W.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Computer Science Department
Illinois Institute of Technology
10 West 31st Street, Chicago, IL 60616

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Cognitive Science Program
Office of Naval Research
800 North Quincy
Arlington, VA 22217-5000

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT
UU

DISTRIBUTION STATEMENT A
Approved for Public Release
Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This grant supported the writing of a book on our experiments in human and computer tutoring and enabled the running of one last experiment with the CIRCSIM-Tutor system, Version 2.9, testing the computer tutor with medical students, using a control group that read a specially designed text.

15. SUBJECT TERMS
intelligent tutoring systems, evaluation, understanding ill-formed input

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Martha W. Evens
					19b. TELEPONE NUMBER *(include area code)*
U	U	U	UU		312-567-5153

Report to the Office of Naval Research on Work Under Grant Number N000140210442 and its Renewal Originally Funded under BAA 02-001 Long-Range Scientific and Technology Program

CONSOLIDATING THE RESULTS OF THE CIRCSIM-TUTOR PROJECT AND FURTHER CONSOLIDATION OF THE RESULTS OF THE CIRCSIM-TUTOR PROJECT

Martha Evens
Department of Computer Science
Illinois Institute of Technology
Chicago, Illinois 60616
evens@iit.edu

Original Starting Date: February 1, 2002

Final Ending Date: September 30, 2003

Total Amount Requested: $4,980

Mail to: Susan F. Chipman, Ph.D.
Office of Naval Research, Code 342
800 N. Quincy Street
Arlington, VA 22217-5660
Phone: 703-696-4318
Fax: 703-696-1212

REPORT ON CONSOLIDATING THE RESULTS OF THE CIRCSIM-TUTOR PROJECT AND FURTHER CONSOLIDATION OF THE RESULTS OF THE CIRCSIM-TUTOR PROJECT

Overview

The purpose of this grant was to support our work on a book for Erlbaum on our experiments in human and computer tutoring and to support one last experiment with the CIRCSIM-Tutor system, Version 2.9, testing the computer tutor with medical students, using a control group that read a specially designed text.

We have made significant progress on the book. We have now written drafts of all twenty-one chapters; some have been reviewed by colleagues, some have not. We still need to make significant additions to Chapters 10, 18, 19, and 21.

At a meeting of ONR Grantees in the tutoring portion of the Cognitive Science Program, our colleagues pointed out a gap in our research results. They suggested that we should compare the learning gains made by students using the CIRCSIM-Tutor system with those made by students reading a carefully edited relevant text. We carried out this experiment in November, 2002. It showed, as we hoped, that 40 students who used CIRCSIM-Tutor for an hour made significantly greater learning gains than the 33 who read a carefully chosen and edited text. Actually 26 of the students in the control group also came and used CIRCSIM-Tutor in the laboratory. A couple of them failed to do the pretest and post-test a second time, however. What is more, over 80% of the students completed all eight problems as opposed to 60% in the experiment in November, 1999.

The system did not crash. It corrected 104 spelling errors without making any miscorrections that we could identify, and did not get caught in any of the confusions that turned up in earlier experiments. The students expressed enthusiasm in the survey. None of the students felt impelled to curse the system. We would like to believe that this was because it is definitely less frustrating to use, but it may just have been due to the number of observers present.

We describe the experimental results below, give a more detailed description of the natural language understanding results, and then summarize some of our other current research.

Experiment with CIRCSIM-Tutor in November 2002

Before we ran the experiment we created a new version, Version 2.9, that corrected many of the problems that appeared in the last major experiment in November, 1999. This version also gave more and better hints and asked a number of open questions.

The changes made by Michael Glass included some errors in the generation grammar that caused the system to generate some ill-formed sentences. He fixed an unintended side effect of the error messages that tell the student what kind of input the system is expecting. In an earlier experiment the system sometimes generated long strings of these messages. Now, if the student does not get the point after two of these error messages in a row, the system tells the student what the answer is and goes on to the next topic. Glass also made some changes to the way the spelling correction system handles phrases that had led to some recognition disasters in 1999. As far as we can tell the system did not miscorrect any spelling in 2002; it did fail to correct "soconstriction" to"vasoconstriction" and "lood volume" to "blood volume."

During the Spring of 2000 Yujian Zhou implemented her four-level student model and used it to improve the classification of the student answers and the hints delivered by the system in response to certain frequent errors. While the system still does not generate as many hints as the experts do, its hinting is much improved.

Reading the transcripts of the machine sessions from Fall, 1999, revealed that the system really short-changed the stronger students by just going on to the next stage or next problem when they filled in a column in the prediction table with correct answers. Expert tutors often ask open questions about the functioning of the baroreceptor reflex at these points or ask the student to make generalizations about the problem-solving process. We had always avoided making the system ask such open questions for fear that it would not be able to parse the answers. We decided that the best way to combine a greater challenge to the student and collect data for extending the parser was to insert such questions into the dialogue, and, without parsing the answer, roll out a "canned" expert answer. This would give the students practice at making explanations, we thought, and still ensure that they saw a correct answer even if the system could not give a tailored critique of that answer. In the event we obtained longer and richer dialogues with a large number of useful answers. Many of them are short answers that we believe the system could parse with only a little work. A number of students realized that the system was not parsing their answers and the result was some interesting testing behavior and some expressions of affect. The next section discusses the language understanding behavior of the system and describes the open questions and the responses received.

We calculated five scores for each pretest and post-test. The pretest scores are precrel (pretest correct relations), prewrel (wrong relations), prepts (points on misconceptions), prepred (pretest predictions), premcq (pretest multiple choice questions – testing transfer to another area in physiology).

	precrel	prewrel	prepts	prepred	premcq	pstcrel	pstwrel	pstpts	pstpreds	pstmcq
C	6.60	2.53	13.95	13.30	2.88	8.50	1.55	18.05	17.03	3.38
E	9.06	1.67	18.52	15.48	3.06	10.45	.82	21.33	17.61	3.61

The post-test asks for the same list of relations and checks on the same misconceptions, asks for predictions on a similar problem and asks multiple choice questions in still another area of physiology. We calculated five scores for each pretest and post-test. The first row in the table, labeled C for Controls (N=33), records the group that took the pre-test, read a text, and took the post-test, all done at home (unsupervised) the weekend before CST laboratory. The second row, labeled E for the Experimental Group (N=40), are students who came to the laboratory, took the pre-test, worked with CIRCSIM-Tutor, then took the post-test and who had not participated in the weekend control group.

For the Controls for each of the five measures, the differences between the means are ALL statistically significant. For the Experimentals for each of the five measures, the differences between the means are ALL statistically significant. Thus we can say that CST "works" but so does the control procedure (reading the text).

We have calculated the difference scores (the gains) for each of the five measures for all students. The differences between the gain scores for Controls and the Experimental are NOT SIGNIFICANT (the one that comes closest to significance is the difference between the gains for correct predictions – the P value is .0587 – with the Experimentals doing better).

Natural Language Understanding in CIRCSIM-Tutor, Version 2.9.

We obtained 66 transcripts from machine sessions on November 10 and 11. There were 40 students who had not been part of the control group doing the reading over the weekend. There were 33 in the control group. So 26 of the students in the control group chose to come to the laboratory as well, but a couple did not do the pretest and post-test. (Note that we did not count M76 where the user logged in and then immediately logged out because of a hardware problem.) We did count M55 and T59 in which the user did precisely one procedure and then logged out.

We will begin with the overall numbers for each session and then discuss the other issues one by one: The Open Questions, the Open Questions that the student actually tried to answer, the Error Turns, the Spelling Correction results and the Number of Procedures Completed. Note that the number of inputs includes blank answers to open questions. The system will not allow blank inputs at other points.

OVERALL NUMBERS FOR INPUTS, OPEN QUESTIONS, SERIOUS ANSWERS TO OPEN QUESTIONS, ETURNS, SPELLING CORRECTIONS, AND PROCEDURES COMPLETED IN NOVEMBER, 2002

Sess.	#S.Inputs	#OpenQS	#OQAns	#Eturns	#SpellErrors	#Procs
M48	47	9	8	1	7	8
M49	62	5	4	2	0	7
M51	85	6	6	2	3	8
M52	22	9	0	1	4	9
M53	32	9	1	0	0	8
M55	18	1	1	0	1	1
M58	55	9	9	1	3	8
M59	99	6	5	4	1	8
M60	62	6	6	2	2	7
M61	93	5	5	2	4	8
M62	39	8	8	0	0	8
M63	63	9	9	3	1	8
M64	31	9	9	0	2	8
M65	35	9	1	2	1	8
M66	35	9	6	2	0	8
M67	99	5	4	12	3	8
M70	68	4	3	7	3	4

M71	50	9	9	3	6	8
M72	62	6	5	4	1	8
M73	45	9	9	0	0	8
M74	17	9	2	0	0	8
M75	52	7	5	4	2	8
M77	46	9	9	0	3	9
M79	38	8	8	0	1	8
M80	48	9	9	0	0	8
M81	99	4	4	4	12	8
M82	45	7	7	2	3	5
M83	61	4	4	5	1	4
M84	54	7	7	0	0	8
M85	51	8	8	1	0	8
T48	11	9	5	0	0	8
T49	41	9	9	0	0	8
T50	25	9	5	4	7	8
T51	49	8	8	1	1	8
T52	64	8	0	7	2	8
T53	39	9	9	4	0	8
T55	61	13	4	4	5	8
T56	28	9	8	2	2	8
T58	54	9	9	1	3	8
T59	9	1	1	0	0	1
T60	38	14	8	1	1	8
T61	21	9	1	2	1	8
T62	20	9	1	1	0	8
T63	36	10	7	4	0	8
T64	36	9	7	2	2	9
T65	49	8	7	0	1	6
T66	46	5	5	1	0	6
T67	86	7	6	5	3	7
T70	16	9	7	1	0	8
T71	51	9	8	2	0	8
T72	61	6	6	4	1	5
T73	31	11	3	1	0	6
T74	35	9	7	2	0	8
T75	22	9	1	1	1	9
T76	42	9	8	0	1	8
T77	47	8	7	3	1	8
T78	55	9	4	5	2	8
T79	10	9	5	0	0	7
T80	50	9	8	0	0	9
T81	22	9	6	0	1	8
T82	19	9	9	0	0	8
T83	30	16	6	1	0	8
T84	22	9	8	1	2	8
T85	59	8	8	4	2	8
T86	46	9	9	0	0	8
T87	36	9	9	2	1	8
TOTAL				130	104	

NUMBERS OF PROCEDURES COMPLETED BY PARTICIPANTS IN NOVEMBER 2002

No. of Procs Completed	No. of Transcripts
1	2
2	0

3	0	
4	1	
5	2	
6	3	
7	4	
8	49	
more than 8	5	(students redid certain procedures)
Total	66	

So 54 (or 81.2%) transcripts include all 8 procedures.

By comparison 60% of the transcripts from Fall 1999 included all eight procedures (21/35).

Procedures Completed in Fall 2002:

M48	1 6 5 4 9 2 3 7	
M49	1 6 5 4 9 2 3	
M51	1 6 5 4 9 2 3 7	
M52	1 6 5 4 9 2 3 7 7	(really)
M53	1 6 5 4 9 2 3 7	
M55	1	
M58	1 6 5 4 9 2 3 7	
M59	1 6 5 4 9 2 3 7	
M60	1 6 5 4 9 2 3	(DR part of 7 and RR predictions)
M61	1 6 5 4 9 2 3 7	
M62	1 2 3 6 5 4 9 7	
M63	1 4 2 3 6 7 5 9	
M64	1 6 5 4 9 2 3 7	
M65	1 6 5 4 9 2 3 7	
M66	1 6 5 4 9 2 3 7	
M67	1 6 5 4 9 2 3 7	
M70	1 6 5 4	(plus DR part of 9 and RR predictions)
M71	1 6 5 4 9 2 3 7	(then returned to 1 and stopped)
M72	1 6 5 4 9 2 3 7	
M73	1 6 5 4 9 2 3 7	
M74	1 6 5 4 9 2 3 7	
M75	1 6 5 4 9 2 3 7	
M77	1 6 5 4 9 2 3 7 1	
M79	1 6 5 4 9 2 3 7	
M80	1 6 5 4 9 2 3 7	
M81	1 6 5 4 9 2 3 7	
M82	1 6 5 4 9 2	
M83	1 6 5 4 9	
M84	1 6 5 4 9 2 3 7	
M85	1 6 5 4 9 2 3 7	
T48	1 6 5 4 9 2 3 7	
T49	1 6 5 4 9 2 3 7	
T50	1 6 5 4 9 2 3 7	
T51	1 6 5 4 9 2 3 2	(plus DR part of 7)
T52	1 6 5 4 9 2 3 7	
T53	1 6 5 4 9 2 3 7	
T55	1 6 5 4 9 2 3 7	
T56	1 6 5 4 9 2 3 7	
T58	1 6 5 4 9 2 3 7	
T59	1	
T60	1 6 5 4 9 2 3 7	

T61	1 6 5 4 9 2 3 7	
T62	1 6 5 4 9 2 3 7	
T63	1 6 5 4 9 2 3 7	
T64	1 6 5 4 9 2 3 7 3	
T65	1 6 5 4 9 2	
T66	1 6 5 4 9 2	
T67	1 6 5 4 9 2 3	(plus DR and RR parts of 7)
T70	1 6 5 4 9 2 3 7	
T71	1 6 5 4 9 2 3 7	
T72	1 6 5 4 9	(plus DR part of 2)
T73	1 6 5 4 9 2	(did DR part of 2 twice, then started 3, did DR)
T74	1 6 5 4 9 2 3 7	(says s/he is T56 but files are different)
T75	1 6 5 4 9 2 3 1 7	
T76	1 6 5 4 9 2 3 7	
T77	1 6 5 4 9 2 3 7	
T78	1 6 5 4 9 2 3 7	
T79	6 5 4 9 2 3 7	(brings up 1 at the end but does not do it)
T80	1 6 5 4 9 2 3 7 1	(plus brings up 6 and does RR and SS)
T81	1 6 5 4 9 2 3 7	
T82	1 6 5 4 9 2 3 7	
T83	1 6 5 4 9 2 3 7	(calls up 1 half way through but does not do it)
T84	1 6 5 4 9 2 3 7	
T85	1 6 5 4 9 2 3 7	
T86	1 6 5 4 9 2 3 7	
T87	1 6 5 4 9 2 3 7	

Note that the order: 1 6 5 4 9 2 3 7
used by all but one student corresponds to doing the procedures
in order down the main menu. The procedure names
that the student sees are:

LIST OF PROCEDURE NAMES AND PROCEDURE NUMBERS

1 DECREASE RA BY 50%
6 INCREASE RV TO 200% OF NORMAL
5 DECREASE IS TO 50%
4 HEMORRHAGE-REMOVE 0.5L
9 HEMORRHAGE-REMOVE 1.0L
2 DENERVATE THE BARORECEPTORS
3 DENERVATE THE BARORECEPTORS AND THEN DECREASE RA BY 50%
7 INCREASE INTRATHORACIC PRESSURE (PIT) TO 2mmHg
8 QUIT THE SYSTEM

My impression from observing students during the Monday session (November 10) is that the students who did fewer than eight procedures were not forced to stop by our time limit but chose to stop because they felt they had done enough. Note that we were paying the students in Fall, 1999, while in Fall, 2002, we paid only the control group. The students used Circsim-Tutor in a regular laboratory.

Note: The system succeeds in breaking out of the eturn pattern and pushing the student on to the next step a number of times (12), including an episode with two eturns in M67 but later in the same session (M67) there are 5 eturns in a row followed by some other problems. What happened here?

Only five of the possible Emessages actually appear in these transcripts:

Evalue:	Please indicate increased, decreased, or unchanged.	18
Emech:	Is the mechanism of control neural or physical?	24
EPT:	Please respond with prediction table parameters.	61
Estage:	Please indicate a stage: DR, RR or SS.	17
Edir:	Please indicate directly or inversely related.	9

The following table contains a list of sessions and the numbers of Eturns in the session. We counted an Eturn as succeeding when the student input was of the right category even if it was incorrect. The last five columns count the eturns of each particular type actually occurring.

SID	#ETurns	#Succeed	#Fail	Evalue	Emech	EPT	Estage	Edir
M48	1	1		1				
M49	2	2			1	1		
M51	2	2			2			
M53	0							
M54	0							
M55	0							
M58	1	1		1				
M59	4	4		1		1	2	
M60	2	2			2			
M61	2	2			1	1		
M62	0							
M63	3	3				1	2	
M64	0							
M65	2	2		2				
M66	2	2		1				1
M67	12	5	7	1	3	8		
M70	7	5	2	1	1	3	2	
M71	3	2	1		1	2		
M72	4	2	2		1	2		1
M73	0							
M74	0							
M75	4	4		1		1	2	
M77	0							
M79	0							
M80	0							
M81	4	4			2	1	1	
M82	2	2			1	1		
M83	5	5			2	3		
M84	0							
M85	1	1					1	
T48	0							
T49	0							
T50	4		4			4		
T51	1	1					1	
T52	7	7		3		1	1	2
T53	4		4			4		
T55	4	3	1		1	2	1	
T56	2	2		1		1		
T58	1	1					1	
T59	0							
T60	1	1				1		
T61	2		2			2		
T62	1	1				1		
T63	4	2	2			2		2
T64	2		2			2		

T65	0							
T66	1	1				1		
T67	5	5		2	1	2		
T70	1	1						1
T71	2	2				1	1	
T72	4	4		1	1		1	1
T73	1	1				1		
T74	2	2			1		1	
T75	1	1				1		
T76	0							
T77	3	3		1	1	1		
T78	5	1	4			4		1
T79	0							
T80	0							
T81	0							
T82	0							
T83	1	1		1				
T84	1	1				1		
T85	4		4			4		
T86	0							
T87	2	2			2			
Total	129	94	35	18	24	61	17	9

The average is almost 2 eturns per session, with the Prediction Table message by far the most common. 94 of the turns were correctly interpreted by the student, 35 were not. These 35 misinterpretations were all made by the 12 students.

```
The following sessions had no eturns: M53.LIS, M55.LIS, M62.LIS,
       M64.LIS, M73.LIS, M74.lis, M77.lis, M79.lis, M80.lis,
       M84.lis, T48.lis, T49.lis, T59.lis, T65.lis,
       T76.lis, T79.lis, T80.LIS, T81.lis, T82.lis, T86.lis
```

The following sessions had two eturns in a row followed by another category error by the student, after which the system changed the subject successfully: M67, M75, T50 twice, T53 twice, T61, T78 twice, and T85 twice.

```
SPELLING CORRECTION:

Number of sessions: 66
Number of sessions with spelling mistakes: 40
Number of sessions with no spelling mistakes: 26
Number of spelling mistakes that the system attempted to correct: 104
```

There were also 53 spelling mistakes that occurred in an answer to open questions - the system did not attempt to analyze this text at all, but we have recorded these errors. Note also: There are 2.6 spelling errors corrected per session where errors occurred - there are 1.58 errors corrected per session.

We also decided to make two extensions to the vocabulary to be considered:

"less" for "decreased/down/-"?
"Ca" or "Calcium" as "a neural mechanism"

The open questions increased the average number of student inputs as did the fact that more people did more procedures.

OPEN QUESTIONS

In conjunction with Khelan Bhatt's study of hedges and affective expressions in the human sessions K52-K76, we determined to look for this kind of behavior in these 66 CIRCSIM-Tutor Sessions. The improvements made by Michael Glass in the spelling correction and parsing and Yujian Zhou's improved hints apparently made the system much less frustrating to use. So did the change in the handling of ETURNS, which led to the student getting a helpful hint in response to the input of a "?" and the system giving the answer after two ETURNS. In any case, the only indications of affect appeared in the answers to open questions. We give the numbers and list all 20 nonblank but nonserious answers to open questions. Many of these are obviously a result of testing the system; so, we think, are some of the ETURNS. The only hedges are two very marginal cases, also in the answers to open questions.

NONSERIOUS ANSWERS TO OPEN QUESTIONS IN November, 2002

See list of all nonblank nonserious answers after table.

SID	#OpenQs	#SeriousAns	#Nonserious	#Blank
M48	9	8	1	1
M49	5	4	1	1
M51	6	6	0	0
M52	9	0	9	5
M53	9	1	8	8
M55	1	1	0	0
M58	9	9	0	0
M59	6	5	1	0
M60	6	6	0	0
M61	5	5	0	0
M62	8	8	0	0
M63	9	9	0	0
M64	9	9	0	0
M65	9	1	8	7
M66	9	6	3	3
M67	5	4	1	0
M70	4	3	1	1
M71	9	9	0	0
M72	6	5	1	1
M73	9	9	0	0
M74	9	2	7	7
M75	7	5	2	2
M77	9	9	0	0
M79	8	8	0	0
M80	9	9	0	0
M81	4	4	0	0
M82	7	7	0	0
M83	4	4	0	0

M84	7	7	0	0
M85	8	8	0	0
T48	9	5	4	1
T49	9	9	0	0
T50	9	5	4	4
T51	8	8	0	0
T52	8	0	8	8
T53	9	9	0	0
T55	13	4	9	9
T56	9	8	1	0
T58	9	9	0	0
T59	1	1	0	0
T60NP	14	8	6	4
T61	9	1	8	8
T62	9	1	8	8
T63	10	7	3	3
T64	9	7	2	2
T65	8	7	1	0
T66	5	5	0	0
T67	7	6	1	1
T70	9	7	2	2
T71	9	8	1	1
T72	6	6	0	0
T73	11	3	8	8
T74	9	7	2	1
T75	9	1	8	8
T76	9	8	1	0
T77	8	7	1	1
T78	9	4	5	5
T79	9	5	4	3
T80	9	8	1	1
T81	9	6	3	0
T82	9	9	0	0
T83	16	6	10	10
T84	9	8	1	1
T85	8	8	0	0
T86	9	9	0	0
T87	9	9	0	0

LIST OF ALL NONBLANK NONSERIOUS ANSWERS

```
M52
T: Why did you predict that IS would not change?
S: 0
M52
T: Why did you predict that IS would increase?
S: 1
M52
T: What does the baroreceptor reflex do?
S: l
M52
T: Can you explain why HR did not change?
S: no
M59
T: What does the baroreceptor reflex do?
S: nothing
```

M65
T: Can you explain why HR did not change?
S: +
M67
T: Why did you enter 'no change' for TPR?
S: BC
T48 Student is testing/teasing?
T: Why did you enter 'no change' for TPR?
S: you know why.
T48
T: Can you explain why HR did not change?
S: yes, i can.
T48
T: Why is MAP still decreased?
S: I don't want to tell you.
T56 Note student types in a tutor response - and gets no recognition -
T: Why did MAP change in the manner that you predicted?
S: If we think about what we mean by DR -- a period of time BEFORE any change in neural activity has occurs, then we can see that heart rate will not change because HR is determined by sympathetic and parasympathetic activity to the hear.
T: Well,the mean aterial pressure changes as a consequence of the DIRECT RESPONSE. The baroreceptor reflex acts to hold MAP as constant as possible, so any change in MAP results in a physiological response which causes MAP to move back in the opposite direction.
T60 NP
T: Why did MAP change in the manner that you predicted?
S: In other words, Nikie knows all...
T60
T: Why did you predict that IS would not change?
S: it
T65
T: CO decreased in DR and increased in RR. Why did you predict that it would decrease in SS?
S: dr
T74
T: Why is MAP still decreased?
S: blalaal
T76
T: Why did you enter 'no change' for TPR?
S: the TPR can
T79
T: Why did you predict that IS would not change?
S: hatever
T81
T: Why is MAP still decreased?
S: asdf
T81
T: What does the baroreceptor reflex do?
S: t
T81
T: Why did you enter 'no change' for TPR?
S: Nimesh said so

POSSIBLE HEDGING IN ANSWERS TO OPEN QUESTIONS:

The following two answers to open questions may possibly be classified as hedges. In both #1 and #2 the student includes a modifier that is not strictly called for: "just yet" in #1 and "9/10" in #2. Can these be called hedges?

1. T: Why did you predict that IS would not change?

Official answer is:

You can think about it this way. Inotropic state is physiologically controlled by the sympathetic nervous system. However, in DR no change in neural activity has occurred yet (the reflex has not started) and so there can be no change in IS.

Student M58 was asked this question during proc 1 (DECREASE RA BY 50%) and answered:

S: because it's a direct response and changing resistance wouldn't affect contractility of the heart just yet

2. T: SV increased in DR and decreased in RR. Why did you predict that it would increase in SS?

The official answer is:
In other words, the change in DR was larger than the compensatory change that occurred in RR. Thus the change in SS is in the same direction as the change in DR.

Student T85 was asked this question in the middle of proc 1 (DECREASE RA BY 50%) and answered:

S: 9/10 times the dr will dominate because the rr can't bring all the way back

Note: NO OTHER EVIDENCE OF HEDGES OF TYPE SEEN IN HUMAN SESSIONS APPEARS ANYWHERE

Note also these spelling correction failures:
M79 CST does not recognize "soconstriction"
M82 "lood volume" is not corrected to "blood volume"

Other Work on the CIRCSIM-Tutor Project during the Last Year

Parsing Long Student Initiatives and Answers. Chung Hee Lee is working on a parser for use if Michael Glass' parser finds too much text it cannot handle. He is currently working on parsing the student initiatives identified by Farhana Shah and the answers to open-ended questions culled from the CIRCSIM-Tutor transcripts from the experiment in November, 2002.

New Knowledge-base Design for CIRCSIM-Tutor and GASP-Tutor. Jay Yusko has reorganized the CIRCSIM-Tutor knowledge base, storing Reva Freeman's rules in a database, and built agents to retrieve separate kinds of information. Bruce Mills is beginning to use this knowledge base in the new Version 3. We believe that this is an important first step in building a more general framework that can be extended to support GASP-Tutor as well.

GASP Vocabulary and Case Frames. Jai Seu and I have identified 139 words and phrases that appear in the four GASP tutoring sessions that are not in CIRCSIM-Tutor. We have built a GASP ontology and we are building case frames for the verbs by hand to use to test the validity of Chung Hee Lee's program to induce case frames automatically. We have also identified the points in GASP sessions where new logic forms are needed and developed half a dozen new ones to express chemical reactions and movement of gases.

Construction of Version 3. Bruce Mills, a Ph.D. student, currently teaching at the College of the Southwest, is building the core of the new Version 3 using Reva Freedman's APE Planner and Jay Yusko's knowledge base of rules.

Analogy in Tutoring. Evelyn Lulis, a Ph.D. student and an Assistant Professor at DePaul University, is investigating the use of analogy in tutoring. We have extracted all the examples of the use of analogy in 75 expert human tutoring sessions. Lulis has now marked them up with information about the base and the target. The markup also records whether or not the tutor asks for an inference based on the analogy and whether the student then got the point and made the correct inference. In case of failure the tutor sometimes repairs the inference and sometimes starts over with a different tutoring strategy. At Dedre Gentner's suggestion Lulis has also marked the analogies up as abstract or concrete. Another important distinction is whether the analogy is based on an earlier student experience with another neural variable or another procedure or whether it is based on prior student knowledge of balloons or of Ohm's Law, say.

Dedre Gentner and Ken Forbus have been extremely helpful and encouraging. We have had three meetings with them and we hope to use the knowledge representation devised by Forbus to map and record the analogies and MacFac to make any necessary repairs. Lulis and Joel Michael are working on a list of analogies to be implemented in Version 3 of the system. Lulis is working with Bruce Mills to implement the turn planner. Then she plans to experiment with extending turn planning to generate analogies in our sessions and to investigate the possibility of repairing analogies misunderstood by the student using Gentner and Forbus' MacFac program.

Hedges and Affective Expressions in Human Tutoring Sessions. There has been a recent surge of interest in trying to understand student affect and hedges. There was considerable discussion of these issues during the SIGDIAL session at ACL 2001 in Pittsburgh. ITS 2002 had a session on emotion/affect - the first such session that we have seen. We have been convinced for some time that we should study the student hedges and the expressions of affect in expert tutoring sessions and consider whether it was desirable for CIRCSIM-Tutor to try to recognize this kind of behavior and devise ways to respond. As part of his MS Thesis work, Khelan Bhatt and Martha Evens have identified 218 hedges (151 hedged answers and 67 hedged initiatives) and 88 affective expressions in the 25 human tutoring sessions conducted in Fall, 1999. All the students hedge but the number of hedges varies widely from 2 in one session to 22 in another. Not all the students express affect, however, and the women students are significantly more likely to express affect than the men. Farhana Shah found that women were significantly more likely to hedge than men in taking the initiative, but we found that men and women did not differ significantly in the proportion of hedged answers.

After the first eight tutoring sessions in 1989, Michael and Rovick decided to stop responding to hedges on the grounds that hedges seemed to say more about the student's preferred style of communication than about the state of the student's knowledge of the subject matter. They did continue to respond in those cases where the student indicated some serious distress or confusion, however. Our results seem to confirm the perceptions of the experts. Although hedged answers are more likely to be in error than answers that are not hedged, more than half of hedged answers are, in fact, correct. If hedging has more to do with interpersonal relationships than with uncertainty then, Khelan suggested, students may hedge differently with novice tutors than with professors. He plans to look at hedging in the novice sessions in the Spring.

We decided that it was time to look for hedging and expressions of affect in the machine tutoring sessions. We found only two examples of hedging in the 66 machine sessions in November, 2002. Both occur in answers to open questions and both are very mild hedges - unnecessary or spurious modifiers "just yet" as opposed to "yet" and "9/10" as opposed to "all". We did, however, find a number of expressions of affect. Students who realized that the system was not parsing their answers to open questions often chose to put in blank answers so that they could look at the canned answer. (In all we received 126 blank answers.) However, we found 20 answers to open questions that display a combination of affective and testing behavior. One student named Nikie answered an open question with "In other words, Nikie knows all ..."

Earlier versions of CIRCSIM-Tutor have also been faced with expressions of affect and been unable to respond appropriately. During our very first trial with students in the early 90's one student typed "abcd" in response to one question and "efgh" in response to the next and the system crashed. The November 1998 sessions garnered a "Kiss my ass" and several cries of "Help." The system responded with a polite error message about the kind of input expected. This was sometimes very helpful, but not very soothing to the frustrated and furious. We expect that students who are using the system at home alone in the middle of the night are generating more curses but unfortunately we have no way to collect the transcripts from these solitary occasions. In November 1999, we saw some more alphabetic runs "jk" and "kl" and a sad comment of "clueless" from lost students. Again, the system responded with an error message. The expert tutors provide help when students do this. Can we devise a way for the system to do this without sounding patronizing?

Publications of the CIRCSIUM-Tutor Project in 2002-2003

Book:

Michael, J.A., Modell, H.I. (2003). *Active learning in secondary and post-secondary science classrooms: a working model for helping the learner to learn.* Upper Saddle River, NJ: Lawrence Erlbaum Associates.

Journal Articles:

Shah, F., Evens, M.W., Michael, J.A., & Rovick, A.A. 2002. Classifying student initiatives and tutor responses in human tutoring keyboard to keyboard tutoring sessions. *Discourse Processes,* 33(1) 23-52.

Michael, J., Rovick, A., Glass, M., Zhou, Y., and Evens, M. (2003). Learning from a computer tutor with natural language capabilities. *Interactive Learning Environments.* 11(3) 233-262.

Book Chapters:

Evens, M. 2002. Thesaural relations in information retrieval. In S.H. Myaeng, C.A. Bean, & R. Green, eds. *The semantics of relationships: An interdisciplinary perspective*. Kluwer, Boston, MA. 143-160.

Evens, M. 2002. Natural language interface for an expert system. In *Encyclopedia of Microcomputers*. Allen Kent & James G. Williams, eds. Marcel Dekker, New York. 225-254. Also in *Encyclopedia of Library and Information Science,* Allen Kent, ed. Marcel Dekker, New York. 228-258.

Papers in Refereed Conferences:

Lee, C.H., Seu, J.H., & Evens, M. 2002. Building an ontology for CIRCSIM-Tutor. *Proc. MAICS 2002*. 161-168.

Lee, C.H., Seu, J.H., & Evens, M. 2002. Automating the construction of case frames for CIRCSIM-Tutor. *Proc. ICAST 2002.* 59-65.

Yusko, J., & Evens, M. 2002. The knowledge collective: Using micro-droids to turn meta data into meta knowledge. *Proc. MAICS 2002.* Chicago, April. 56-60.

Lulis, E., & Evens, M.W. (2003). The Use of Analogies in Human Tutoring Sessions. *Proc. AAAI Symposium on Natural Language Generation in Spoken and Written Dialogue.* March, Stanford, CA. 94-96.

Lee, C.H., & Evens, M.W. (2003). Interleaved Syntactic and Semantic Processing for CIRCSIM-Tutor Dialogues. *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference MAICS'03,* pp. 69-73. Cincinnati, OH.

Zhao, J., Kim, J.H., & Evens, M. (2003). Comparison of Student Initiatives in Keyboard-to-Keyboard and Face-to-Face Tutoring Sessions. *Proceedings of the Midwest Artificial Intelligence and Cognitive Science Conference, MAICS'03.* pp. 178-183. Cincinnati, OH.

Jeong, I., Evens, M.W., & Kim, Y. (2003). Automatic Knowledge Acquisition Using Concept Map Generation. *Proc. ICAI'03.* June, Las Vegas, Nevada.

Lulis, E., Evens, M.W., & Michael, J.A. (2003). Representation of Analogies in Tutoring Sessions. *Proc. of the Second IASTED Conference on Information and Knowlege Sharing*. pp. 88-93. Scottsdale, AZ.

Mills, B., & Evens, M.W. to appear. Implementing the Directed Line of Reasoning with the Atlas Planning Environment. *Proceedings of ITCC, 2004.*